Video Compression Based on Spatio-Temporal Resolution Adaptation

Mariana Afonso, *Student Member, IEEE*, Fan Zhang, *Member, IEEE*, and David R. Bull, *Fellow, IEEE*

***Abstract*— A video compression framework based on spatio-temporal resolution adaptation (ViSTRA) is proposed, which dynamically resamples the input video spatially and temporally during encoding, based on a quantisation-resolution decision, and reconstructs the full resolution video at the decoder. Temporal upsampling is performed using frame repetition, whereas a convolutional neural network super-resolution model is employed for spatial resolution upsampling. ViSTRA has been integrated into the high efficiency video coding reference software (HM 16.14). Experimental results verified via an international challenge show significant improvements, with BD-rate gains of 15% based on PSNR and an average MOS difference of 0.5 based on subjective visual quality tests.**

***Index Terms*— Video compression, spatial resolution adaptation, temporal resolution adaptation, perceptual video compression, CNN-based super-resolution.**

I. Introduction

With the ever increasing demand for more immersive visual experiences, video content providers have been extending the video parameter space by using higher spatial resolutions, frame rates and dynamic ranges. This dramatically increases the bitrate required to store and distribute video content, challenging current bandwidth limitations and demanding greater compression efficiency than offered by the current generation of video codecs.

Previous work has shown that the optimal parameters for video representation with respect to perceptual quality is highly content dependent [1], [2]. Thus, by dynamically predicting these parameters, bitrates could be significantly reduced while maintaining equivalent perceptual video quality. In this context, several authors have proposed reducing spatial resolution for low bitrate encoding [3], [4], but lack a reliable adaptation technique. Others have developed prediction models [5], [6] or have introduced the resolution adaptation as one of the rate-distortion optimized modes at a block level (CTU) [7] but apply them for H.264 or intra coding only. Regarding temporal adaptation, a few methods for frame rate selection have been proposed in [8] and [9]. However these have not been fully integrated with video compression algorithms. Moreover, the reconstructed video quality depends highly on the video resampling technique applied. Previous spatial resolution adaptation approaches have mostly employed linear filters, such as bicubic, for the reconstruction of full resolution video frames. However, in recent years, CNN-based super-resolution techniques [10], [11] have become popular in the field of computer vision due to the improved reconstruction quality. Such machine learning-based approaches have however not been fully explored for video compression.

Inspired by our previous work on quality assessment [1], [2], [12], [13] and spatial resolution adaptation for intra coding [14], we propose a spatio-temporal resolution adaptation framework for video compression, ViSTRA, which dynamically predicts the optimal spatial and temporal resolutions for the input video during encoding and attempts to reconstruct the full resolution video at the decoder.

The main contributions of our paper include:

- The integration of both spatial and temporal adaptation into a single framework;
- A Quantization-Resolution Optimization (QRO) module which applies perceptual quality metrics and machine learning techniques to generate reliable resolution adaptation decisions;
- The employment of a CNN-based super resolution model to reconstruct full spatial resolution content, trained specifically for compressed content;
- The integration of the ViSTRA framework with HEVC reference software (HM 16.14).

The experimental results presented here, are based on test sequences used in the Video Compression Grand Challenge at IEEE ICIP 2017 [15]. These show substantial coding gains, on average 14.5% BD-rate (PSNR) and 0.52 average MOS difference (from independent subjective test), compared to the original HEVC anchor codec (HM 16.14).

The remainder of this paper is organised as follows: Section II describes the proposed framework; Section III provides more detail into the design of the QRO module; Section IV describes the employed methods for spatial and temporal resolution resampling; Section V presents and discusses the experimental design and results, and finally section VI provides conclusions and ideas for future work.

II. Proposed Framework

The proposed framework, shown in Fig. 1, integrates spatio-temporal adaptation with video encoding in order to maximize rate-quality performance. As the first step, video frames of the full resolution video are processed by the QRO module,

Manuscript received May 3, 2018; revised August 7, 2018; accepted October 23, 2018. Date of publication October 31, 2018; date of current version January 7, 2019. This work was supported in part by FP7 Marie Skodowska-Curie under Actions Grant 608231-PROVISION ITN, in part by Google Faculty Research Awards 2016, and in part by EPSRC under Grant EP/M000885/1. This paper was recommended by Associate Editor M. Wien. *(Corresponding author: Mariana Afonso.)*

The authors are with the Visual Information Laboratory, University of Bristol, Bristol BS8 1TH, U.K. (e-mail: mariana.afonso@bristol.ac.uk; fan.zhang@bristol.ac.uk; dave.bull@bristol.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2878952

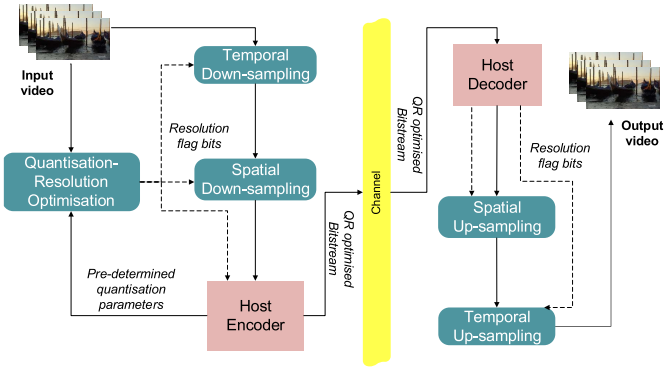

Fig. 1. Diagram of the proposed resolution adaptation framework for video compression.

which is responsible for predicting the suitability for both spatial and temporal adaptation, given the content of the video and the input quantisation parameter (QP). Two decisions are made: one for spatial and one for temporal adaptation. These decisions then control the modules that apply spatial and temporal downsampling, respectively. The adaptation is signaled in the bitstream using flag bits. The input of the host encoder is therefore the resolution optimized video. At the decoder, the flag bits are extracted from the bitstream and resolution resampled video frames are decoded by the host decoder. Finally, video frames are spatially and temporarily upsampled to the original resolution for displaying based on the resampling decisions from the encoder.

Temporal adaptation decisions are made between 2 frames and require one flag bit per frame to be added to the bitstream. In contrast, Spatial adaptation decisions are made for each Group-of-Pictures (GOP) requiring one flag bit per GOP. If two subsequent GOPs are found to contain different spatial decisions, e.g., the resolution of the second GOP is half the resolution of the first GOP, a split is introduced at that point and two separate bitstreams are encoded. This procedure is required due to the fact that the HEVC does not inherently support the encoding of different spatial resolutions.

III. Quantisation-Resolution Optimization

Before compression, full resolution video frames are sent to the QRO module, where temporal and spatial resolutions are subsequently optimized according to the initial quatisation parameter and video content.

A. Temporal Decisions

Temporal resolution optimization employs a frame rate dependent quality metric, FRQM [12], to assess the perceptual quality difference between a temporally downsampled video frame and its full frame rate original. FRQM is a state-of-the-art quality metric that provides the best performance for perceptual evaluation of frame-rate reduction artifacts. Temporal downsampling is achieved using frame averaging. If the resulting FRQM score is higher than a pre-determined threshold, TH, the temporally downsampled video frame is encoded in place of the original versions. In this work, only a ratio of 2 was used for downsampling, and each decision is made within a time window of 2 frames. The temporal resolution flag, $\mathrm{TR}_{\mathrm{flag}}$, which indicates if temporal resolution adaptation is performed, is computed as follows:

$$\mathrm{TR}_{\mathrm{flag}} = \begin{cases} 1, & \text{if FRQM} > \mathrm{TH} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $\mathrm{TH} = 48$, which is the FRQM score that corresponds to a DMOS of 10 (from 0 to 100) based on the relationship between FRQM and DMOS predicted in [12]. This ensures that there are no significant perceptual differences between the temporally resampled video and the original video. Based on [1] and [12], we expect that temporal resampling to be more beneficial for videos with higher frame rates (from 120 fps and higher).

B. Spatial Decisions

The spatial resolution decision module employ a learning-based approach using low level spatio-temporal features from the uncompressed video frames, following a similar methodology as the one applied in [14]. These features are computed on the Y (luma) channel only and are used to predict a Quantization Parameter (QP) threshold, $\mathrm{QP}_{\mathrm{thres}}$, at which encoding the input video at a lower resolution will produce higher rate-quality performance compared to encoding at the original resolution. In this work, a single ratio of 2 is used for resampling.

In our previous publication on intra coding [14], we proposed a module that predicted QP thresholds based on the PSNR of each frame downsampled and upsampled using a Lanczos filter (kernel size 3), the resampling PSNR. However, for random access configuration, features that measure temporal correlations are also required. Therefore, in this paper, in addition to PSNR, two spatio-temporal features are calculated for consecutive frames, Normalized Cross-Correlation (NCC) and Temporal Coherence (TC), which are defined in [16]. In total, 4 features are used for prediction forming the feature vector $\mathbf{K}$ as follows:

$$\mathbf{K} = [\mathrm{PSNR}_{\mathrm{r}}, \mathrm{NCC}_{\mathrm{skewness}}, \mathrm{TC}_{\mathrm{kurtosis}}, \mathrm{TC}_{\mathrm{skewness}}] \tag{2}$$

where $\mathrm{PSNR}_{\mathrm{r}}$ is the resampling PSNR, $\mathrm{NCC}_{\mathrm{skewness}}$ is the skewness of the NCC, $\mathrm{TC}_{\mathrm{kurtosis}}$ is the kurtosis of the TC and $\mathrm{TC}_{\mathrm{skewness}}$ is the skewness of the TC. This particular set of features was obtained by forward feature selection using 5-fold cross-validation.

We represent the relationship between the spatio-temporal features, $\mathbf{K}$ of the video content and the spatial QP threshold, $\mathrm{QP}_{\mathrm{thres}}$, using linear regression given by:

$$\mathrm{QP}_{\mathrm{thres}} = \mathbf{W} \cdot \mathbf{K}' + \beta \tag{3}$$

where $\mathbf{W}$ and β are the fitting parameter.

The features are computed for a training dataset, consisting of 57 temporally cropped UHD videos from the Harmonic Inc video database [17]. After calculating the feature vector and true QP thresholds (based on multiple encodes with HM 16.14) for all sequences, the fitted linear regression parameters are given by $\mathbf{W} = [-0.62, 1.94, -0.58, -3.87]$ and $\beta = 63.7$.

The Root Mean Square Error (RMSE) of the fit on the training data is 3.26 (QP values).

Finally, for the test sequences, a simple comparison between the predicted QP threshold, QP_{thres} and the input QP used for encoding, QP_{in} is performed, which determines the spatial resampling decision SR_{flag}, as shown below:

$$SR_{flag} = \begin{cases} 1, & \text{if } QP_{in} > QP_{thres} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

In addition, if resolution downsampling is applied, the base QP used to encode the low resolution video is reduced by a default value of 6, following the analysis in [7] and [14], in order to achieve a similar bitrate as would be achieved if no resampling was performed.[1]

IV. Spatial and Temporal Resampling

This section introduces the methods used for temporal and spatial resampling. In ViSTRA, these resampling methods are applied separately (see Fig. 1).

A. Temporal Resampling

Temporal downsampling is achieved using frame averaging, a technique commonly applied for adapting video frame rates [1], [18]. Compared to dropping frames, this has been shown to provide more perceptually pleasing artifacts and better frame rate up-conversion. Moreover, nearest-neighbor interpolation is used for upsampling which emulates a hold-type display (e.g. LCD) [1].

B. Spatial Resampling

In our previous work [14] we have proposed using Lanczos3 filters for both downsampling and upsampling. However, in order to further improve reconstruction performance for upsampling, ViSTRA employs a deep CNN.

Although previously proposed CNN-based methods for super-resolution, including SRCNN [10] and VDSR [11], provide exceptional performance, they are not directly applicable to video coding because their models were training with uncompressed images. We therefore use the CNN architecture of VDSR (Fig. 2) and retrained it for HEVC compressed video. This architecture contains 20 convolutional layers with 64 3×3 filters followed by ReLU activation functions and applies residual learning.

The same dataset used for training the QRO module (Section III) was also used for training the CNN. The training mechanism works as follows: first, each sequence is downsampled using a Lanczos3 filter with a downsampling ratio of 2. Then the low resolution versions are encoded and decoded using HEVC (HM 16.14), random access main configuration using 4 different QP values (22, 27, 32 and 37). The reconstructed videos are then upsampled using the same filter. The frames that result from this process are then used as training inputs to the CNN with the output targets being the original uncompressed frames.

[1]Note that the Lagrange multipliers used in the encoder's Rate-Distortion Optimization (RDO) process follows the same relationship as described in the HEVC reference software, HM 16.14.

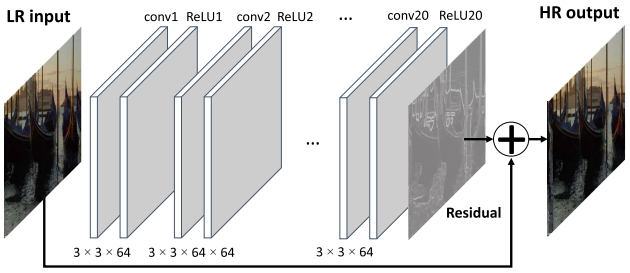

Fig. 2. Network architecture applied to spatial resolution upscaling.

We select a subset of the frames of the resulting video, split them into 41×41 pixels blocks (the size of the receptive field of the CNN) and choose 400 blocks randomly. In addition, in order to provide more generalization for the CNN, data augmentation is applied in the form of block rotation. Therefore, a total of approximately 4 million blocks were used to train the CNN. Finally, training was performed using CAFFE [19] and the following training parameters: Adam optimization [20], batch size of 64, learning rate of 1e-4 (fixed), Weight decay of 1e-4 and 10 epoch.

V. Experimental Results

The proposed framework was integrated into HEVC test model HM 16.14 and was submitted to the Grand Challenge on Video Compression Technology at the International Conference on Image Processing (ICIP) 2017 [15]. The aim of this Challenge was to identify technologies that provide significant improvement beyond the current state of the art in video compression. It is important to note that the subjective results presented have been obtained independently by the organizers of the challenge.

The test dataset is composed of 9 sequences, 4 HD (1920×1080) and 5 cropped UHD (2560×1600), obtained from the JVET (Joint Video Exploration Team) UHD test set [21] and the BVI Texture database [22]. In addition, the organizers provided 4 target rate points per sequence and the respective HEVC (HM 16.14) anchors per rate point. These were selected mainly in order to provide low quality anchors which could be perceptually improved by the submissions. In order to meet the target bit rates for the test sequences, the QP values were iteratively adjusted until the output bitrates became sufficiently close to the target bitrates. A sample frame and target rate points for each sequence can be found in Fig. 3.

The test results are based on PSNR, Video Multimethod Assessment Fusion (VMAF) [23] and subjective tests. For the subjective tests, a single-stimulus methodology with 28 participants conforming to the home environment conditions outlined in BT.500-13 [24], was conducted using the submissions received and the anchors. TABLE I shows the Bjøntegaard measurements [25] on PSNR, VMAF and subjective MOS for the proposed method against the HEVC anchor. In addition, Fig. 4 presents the rate-quality curves for two test sequences, LampLeaves (S03) and CatRobot (S05). The BD-MOS values were computed by following the procedure in [26].

Fig. 3. Test sequences and target bitrates used for experimental results: proposed for the Grand Challenge on Video Compression Technology at ICIP 2017. All sequences are 60 fps except for ParkRunning which is 50 fps.

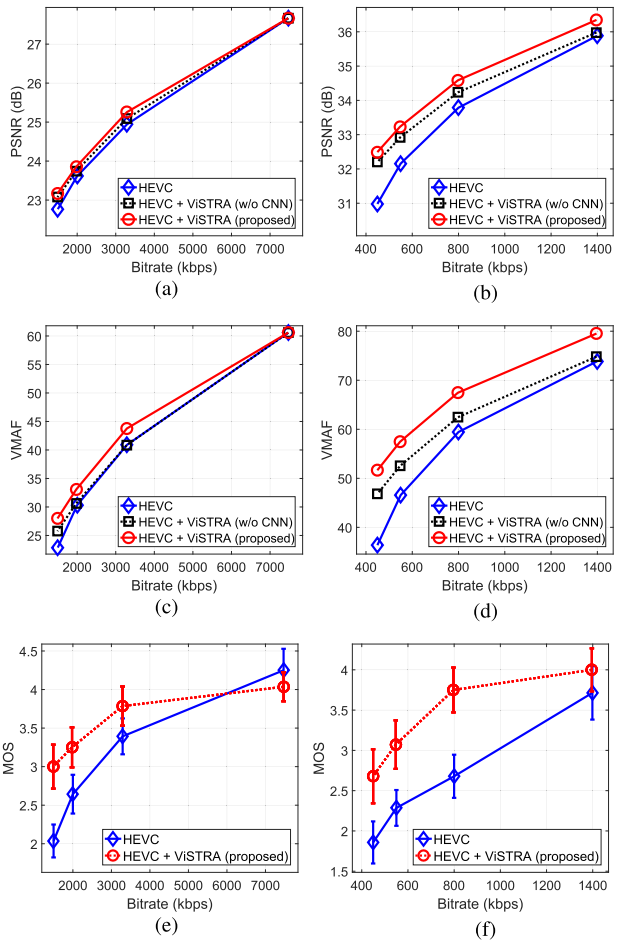

Fig. 4. Rate-quality curves comparing the anchor HEVC (HM 16.14), ViSTRA without the CNN and the proposed approach. Subjective tests (MOS) were only performed using the anchor and the proposed methods. The error bars represent 95% confidence intervals. (a) LampLeaves-S03 (PSNR). (b) CatRobot-S05 (PSNR). (c) LampLeaves-S03 (VMAF). (d) CatRobot-S05 (VMAF). (e) LampLeaves-S03 (MOS). (f) CatRobot-S05 (MOS).

It is noted that significant improvements have been achieved for the proposed method over the anchor codec, with an average of 14.5% BD-rate gains (using PSNR) and 0.55 BD-PSNR. When using VMAF, which correlate better with subjective quality [27], the results are more pronounced with an average of 21.2% BD-rate and 6.1 BD-VMAF. Finally, the subjective tests confirm the perceptual quality gains achieved by the proposed framework, with an average BD-MOS of 0.52.

TABLE I
EXPERIMENTAL RESULTS OF THE PROPOSED METHOD COMPARED TO HEVC HM 16.14 ANCHOR

Sequence	BD-rate (PSNR) [%]	BD-PSNR [dB]	BD-rate (VMAF) [%]	BD-VMAF	BD-rate (MOS)	BD-MOS
S01	-12.7	0.33	-19.2	4.3	-18.1	0.26
S02	-1.4	0.13	-11.2	2.8	-41.4	0.49
S03	-9.9	0.28	-11.9	2.8	-26.4	0.32
S04	-17.5	0.52	-23.5	5.8	-29.6	0.50
Avg. HD	**-10.4**	**0.32**	**-16.5**	**3.9**	**-28.9**	**0.40**
S05	-19.6	0.87	-26.3	8.9	-44.0	0.81
S06	-14.9	0.51	-21.5	6.8	-34.8	0.66
S07	-16.9	0.80	-24.4	8.2	-25.4	0.42
S08	-20.9	1.04	-29.7	9.4	-46.5	0.77
S09	-16.4	0.49	-23.1	5.7	-36.0	0.44
Avg. UHD	**-17.7**	**0.74**	**-25.0**	**7.8**	**-37.3**	**0.62**
Avg. total	**-14.5**	**0.55**	**-21.2**	**6.1**	**-33.6**	**0.52**

The results show that ViSTRA achieves more significant gains for higher spatial resolutions, 17.7% BD-rate (PSNR) for 2560 × 1600 test sequences compared to 10.4% for 1920 × 1080. This is due to the increased spatial redundancy for higher resolutions, which means that less information may be lost by applying the downsampling process.

Figure 4 also compares the rate-quality performance of ViSTRA without the use of the CNN at the decoder for the two example sequences. On average, the use of the CNN increases the quality of the reconstructed frames by 0.19 dB and 3.5 VMAF values, which is reflected in an increase of BD-rate by 6.0% based on PSNR and 14.1% based on VMAF.

For the majority of sequences and rate points tested, only spatial resampling is invoked, except for TreeWills. This is due to the fact that the highest frame rate used for the test sequences is 60 fps and temporal resampling is mostly beneficial at higher frame rates or for sequences containing slow motion, which is the case for TreeWills. A separate analysis shows that spatial adaption is responsible for 12.9 % of the total BD-rate gains based on PSNR for this sequence. Additionally, no adaptation was applied for only one rate point, LampLeaves at 7500 kbps, in which the original resolution was encoded (see Fig. 4 (a)).

In relation to the complexity of the proposed approach, the average encoding time is reduced to 0.58 times that of the HM 16.14 encoder. This is due to the fact that ViSTRA allows the encoding of reduced spacial and temporal resolution videos, which decreases the encoding time significantly. However, the average decoding time of ViSTRA is on average 61 times that of HM, due to the application of the CNN for spatial resolution upscaling. These values were obtained on a shared cluster from the University of Bristol which contains SandyBridge CPUs with 16 cores, 2.6 GHz clock speed and 4GB memory each. The decoding jobs were run in GPU nodes with NVIDIA K20.

VI. Conclusion

This paper proposes a spatio-temporal resolution adaptation framework for video coding, ViSTRA, which optimally resamples input video frames during encoding and reconstructs the full resolution video frames at the decoder. We propose a quantization-resolution module which computes features from the original uncompressed input video frames and determines the optimal spatial and temporal resolution at which to encode them. At the decoder, we apply frame repetition and a Convolutional Neural Network (CNN) for temporal and spatial resolution upscaling, respectively. This framework has been integrated into HEVC test model HM 16.14 and extensive experimental results were conducted using objective quality metrics and subjective tests. These show that significant coding gains can be achieve by applying the proposed framework for video coding. Future work will focus on improving the performance and reducing the complexity of the CNN for spatial-temporal resampling and on testing more immersive video formats including 4K resolution and 120 fps sequences.

VII. Acknowledgment

The authors would like to thank the Harmonic Inc. for permission to use their 4K sequences for research and the committee of the Grand Challenge in Video Compression Technology at ICIP 2017 for permission to use the subjective test results.

References

[1] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3407–3411.

[2] A. Mackin, M. Afonso, F. Zhang, and D. R. Bull, "A study of subjective video quality at various spatial resolutions," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 2830–2834.

[3] M. Shen, P. Xue, and C. Wang, "Down-sampling based video coding using super-resolution technique," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 6, pp. 755–765, Jun. 2011.

[4] G. Georgis, G. Lentaris, and D. Reisis, "Reduced complexity superresolution for low-bitrate video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 2, pp. 332–345, Feb. 2016.

[5] R.-J. Wang, C.-W. Huang, and P.-C. Chang, "Adaptive downsampling video coding with spatially scalable rate-distortion modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1957–1968, Nov. 2014.

[6] J. Dong and Y. Ye, "Adaptive downsampling for high-definition video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 480–488, Mar. 2014.

[7] Y. Li *et al.*, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2316–2330, Sep. 2018.

[8] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, "Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 671–682, May 2012.

[9] Q. Huang *et al.*, "Perceptual quality driven frame-rate selection (PQD-FRS) for high-frame-rate video," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 640–653, Sep. 2016.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[11] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[12] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 300–304.

[13] A. Mackin, M. Afonso, F. Zhang, and D. R. Bull, "SRQM: A video quality metric for spatial resolution adaptation," in *Proc. Picture Coding Symp. (PCS)*, 2018, pp. 283–287.

[14] M. Afonso, F. Zhang, A. Katsenou, D. Agrafiotis, and D. Bull, "Low complexity video coding based on spatial resolution adaptation," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 3011–3015.

[15] *Grand Challenge at ICIP 2017: Video Compression Technology*. Accessed: Jan. 3, 2017. [Online]. Available: http://www.provision-itn.eu/grand-challenge-videocompression-icip2017.html

[16] A. V. Katsenou, M. Afonso, D. Agrafiotis, and D. R. Bull, "Predicting video rate-distortion curves using textural features," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2016, pp. 1–5.

[17] *Harmonic Inc 4K Demo Footage*. Accessed: May 1, 2017. [Online]. Available: https://www.harmonicinc.com/4k-demo-footage-download/

[18] M. Armstrong, D. Flynn, M. Hammond, S. Jolly, and R. Salmon, "High frame-rate television," BBC Res., White Paper 169, Sep. 2008.

[19] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[20] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[21] *JVET Common Test Conditions and Software Reference Configurations*, document JVET-B1010, 2nd JVET Meeting, JVET, San Diego, CA, USA, Feb. 2016.

[22] M. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "A video texture database for perceptual compression and quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 2781–2785.

[23] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Netflix, Los Gatos, CA, USA, The Netflix Tech Blog, 2016. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-videoquality-metric-653f208b9652

[24] International Telecommunications Union—Radiocommunication (ITU-R), Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[25] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," document VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2001.

[26] P. Hanhart and T. Ebrahimi, "Calculation of average coding efficiency based on subjective quality scores," *J. Vis. Commun. Image Represent.*, vol. 25, no. 3, pp. 555–564, Apr. 2014.

[27] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2620–2630, Oct. 2018.

Mariana Afonso received the B.S./M.S. degree in electrical and computers engineering from the University of Porto in 2015. She is currently pursuing the Ph.D. degree in video compression at the University of Bristol, U.K. During her Ph.D., from 2015 to 2017, she was a part of the PROVISION ITN European Commission's FP7 Project, a network of leading academic and industrial organizations in Europe, working on the perceptual video coding. She also completed a secondment with the Video Algorithms Team at Netflix, Los Gatos, CA, USA, in 2017. Her research interests include video compression, video quality assessment, and machine learning.

Fan Zhang received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University in 2005 and 2008, respectively, and the Ph.D. degree from the University of Bristol in 2012. He is currently a Research Assistant with the Visual Information Laboratory, Department of Electrical and Electronic Engineering, University of Bristol, working on projects related to perceptual video compression. His research interests include perceptual video compression, video quality assessment and immersive video formats, including HDR and HFR.

David R. Bull received the B.Sc. degree from the University of Exeter, Exeter, U.K., in 1980, the M.Sc. degree from The University of Manchester, Manchester, U.K., in 1983, and the Ph.D. degree from the University of Cardiff, Cardiff, U.K., in 1988. He was previously a Systems Engineer with Rolls Royce, Bristol, U.K., and also a Lecturer at the University of Wales, Cardiff. He joined the University of Bristol in 1993, where he is currently the Chair of Signal Processing and the Director of the Bristol Vision Institute. In 2001, he co-founded a university spin-off company, ProVision Communication Technologies Ltd., specializing in wireless video technology. He has authored over 450 papers on the topics of image and video communications and analysis for wireless, Internet, and broadcast applications, together with numerous patents, several of which have been exploited commercially. He is a fellow of the Institution of Engineering and Technology. He has received two IEE Premium awards for his work. He has authored three books, and has delivered numerous invited/keynote lectures and tutorials. He is a fellow of the Insititution of Engineering and Technology.